DOCUMENT RESUME

ED 343 929 TM 018 077

AUTHOR Klein, Thomas W.
TITLE Procedures for Scaling the 1990 Edition of the Nevada Proficiency Examinations in Reading and Mathematics.
PUB DATE 12 Jan 91
NOTE 5p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Achievement Tests; Cutting Scores; Grade 11; High Schools; *High School Students; *Item Analysis; Item Response Theory; *Mathematics Tests; Pilot Projects; *Reading Tests; *Scaling; Standardized Tests; *State Programs; State Standards; Test Construction; Testing Programs; Test Items; Test Results; Test Validity
IDENTIFIERS *Nevada High School Proficiency Examinations

ABSTRACT

Steps involved in the item analysis and scaling of the 1990 edition of Forms A and B of the Nevada High School Proficiency Examinations (NHSPEs) are described. Pilot tests of Forms A and B of the 47-item reading and 45-item mathematics tests were each administered to random samples of more than 600 eleventh-grade students. A computer program was developed to calculate the classical item statistics and tabulate the proportion of items answered correctly by students in each quintile. These proportions provided information similar to that attainable from item characteristic curves in item response theory (IRT) analysis. Analysis determined that items of Forms A and B of the mathematics test were acceptable. One item on Form A of the reading test was deleted, and nine other items on Forms A and B of the reading test were rewritten to improve their ability to discriminate. The BILOG computer program was then used to confirm these results, scaling the forms using the three-parameter IRT model. These scalings were used to set the cutting scores. Cutting scores were compared to results from the same students from the previous edition of the NHSPEs. There was good agreement between results of the two tests. The Nevada Department of Education is recommending that the passing standard be raised with the expectation that about 16% of students taking the NHSPEs the first time will fail. Two tables present study findings. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

Procedures for Scaling the 1990 Edition of the Nevada Proficiency Examinations in Reading and Mathematics

ED343929

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Thomas W. Klein, Ph.D., Director
Nevada Proficiency Examination Program

January 12, 1991

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
KEVIN CROWE
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Introduction

The 1990 edition of the High School Proficiency Examinations in reading and mathematics were developed by the Nevada Department of Education in collaboration with district and school representatives. Forms A and B of each examination were further refined by CTB/McGraw Hill, which produced the camera ready copy for the pilot tests. The pilot test in mathematics consisted of 45 items representing five skill areas and in reading consisted of 47 items representing eight skill areas. The items for the mathematics test were those submitted to CTB by the test development group. Since copyright releases were not available for several of the reading passages that formed the basis for the reading tests, CTB supplemented those materials with passages and items from their item bank in producing the pilot tests.

The pilot tests were administered to random samples of 11th grade students, statewide, within one week of the administration of the High School Proficiency Examinations in February, 1990. More than 600 students took the pilot test of each form of each examination. These samples were non-overlapping so that each student involved in the pilot test took only one form of one test.

The scaling of the 1990 editions of the reading and mathematics tests was conducted in two stages. First, item analysis was carried out for each form of each test using the data from the pilot test. Weak items were identified and modified to increase their validity and contribution to the total score on the test. The second stage involved the scaling of the pilot tests, deleting those items that had been rewritten and using the three-parameter item response model, setting the mean to 500 and the standard deviation to 100. These results provided the basis for setting the cut scores for those 12th grade and adult students who took Form A of the 1990 Edition in October, 1990. This process was repeated using the data from those students who took the examinations for the first time in October, 1990, in order to conduct the cut score analysis for the final tests containing the rewritten items.

The purpose of this communication is to document the steps involved in the item analysis and scaling of the 1990 Edition of Forms A and B of the High School Proficiency Examinations in reading and mathematics.

TM018077

BEST COPY AVAILABLE

Item Analysis of the Pilot Tests

The sample size for the item analysis and the median raw score for each form of the reading and mathematics pilot tests are indicated in Table 1.

It was intended that item analysis be carried out with BILOG-PC, a PC based program implementing item scaling and test scoring using the Item Response Theory (IRT) model. Delays in obtaining this program from the vendor necessitated the development of a program to calculate the classical item statistics and tabulate the proportion of items answered correctly by students in each quintile. These proportions provide information similar to that obtainable from the item characteristic curves in IRT analysis.

Analysis using the internally developed program indicated that the items of both Mathematics Form A and Form B were acceptable. A few items in each form had biserial correlations (r_{bis}) with the total score that were less than .5, but these items measured skills important to the test and were retained. Despite inclusion of these items with weaker correlations, the internal consistency of the tests as measured by Kuder-Richardson formula 20 (KR-20) was quite good. The minimum r_{bis}, the percentage of items with biserial correlations greater than .5, and the reliability coefficient for each form of each test are also included in Table 1.

Both forms of the reading tests contained items which either did not discriminate well between the more and less able students or had only marginal correlations with the total score. Six items in Form A and four items in Form B fell in this category. Of these, one item in Form A was negatively correlated with the total score. Based on these results, one item was deleted from Form A and the remainder of the weak items in the two forms of the reading test were rewritten in an attempt to improve their ability to discriminate and improve their relationship with the whole.

Table 1. Results of the initial analysis of all original items from the two forms of the 1990 Edition of the High School Proficiency Examinations in reading and mathematics.

Test	Form	N	Median Raw Score	Minimum r_{bis}	Percent $r_{bis}>.5$	K-R 20
Reading (47 items)	A	686	34	-.110	76.6	.926
	B	687	36	.142	83.0	.924
Mathematics (45 items)	A	661	34	.380	80.0	.902
	B	682	36	.221	88.9	.917

Three-Parameter IRT Scaling

When BILOG-PC was finally received, the pilot tests were run to confirm the item analysis from the local program. Following this confirmation, the two forms of the mathematics

tests were scaled using the three-parameter IRT model with a mean of 500 and a standard deviation of 100. The two forms of the reading test were scaled in the same manner, omitting those items that had been rewritten for the final form of the test. These scalings of Form A were used to set the cut scores of 370 for those 12th grade and adult students who took the examination in October, 1990. Thus, only 41 items of Reading Form A were scored for this group.

In three-parameter IRT scoring, each item contributes differentially to the total score as a function of the item's ability to discriminate between the more and less able students, its difficulty, and the extent to which it is estimated that students guessed on the item. Thus, there is no one-to-one correspondence between raw and scaled scores. The same raw score can yield somewhat different scaled scores depending on the characteristics of the specific items answered correctly.

Since we currently score proficiency examinations on the basis of number correct, and are not yet set up to apply IRT scoring on the scale required by the Proficiency Examination Program, a link needed to be developed between the raw score, or number of items answered correctly, and the IRT scale. To accomplish this, the mean IRT scaled score was calculated for each possible raw score. These values were then plotted, and a smooth curve was drawn among these points to remove irregularities which might have been caused by small sample sizes for particular values of the raw score. These curves were then used to determine the raw score equivalent of the scaled score cutpoint, and the scaled score equivalents of each raw score reported in the score distributions for schools, districts, and the state.

To test the validity of these cutpoints, it was first determined which of the students who had taken the pilot test would have passed the test and which would have failed, using the 370 cutpoint. The distributions of the scores on the High School Proficiency Examination, that these students had taken about a week before the pilot test, was then produced for these two groups for each form of each test. Inspection of the distributions of HSPE scores revealed good agreement between the results of the two tests.

The Recommendation for Raising the Standard for Passing Reading and Mathematics

One of the few documents in our files which deals with the standard setting for the 1984 and 1986 editions of the High School Proficiency Examinations expresses the intent to reduce the passing rate in reading from approximately ninety-five percent to approximately ninety percent and to maintain the passing rate for mathematics at approximately ninety percent, the passing rate for the 1982 examination. The passing rates for grade 11, presented in Table 2, for the period 1986-1989 indicate that the expected rates from the IRT scaling were overestimates of the actual rates. That is, more students passed than expected and the average rate for passing these examinations has been almost ninety-five percent.

Table 2. Number of eleventh grade students taking and percent passing the mathematics and reading tests of the High School Proficiency Examinations on their first attempt for the period 1985-86 through 1988-89.

Academic Year	Reading		Mathematics	
	Number	% Passing	Number	% Passing
1985-86		98		95
1986-87	10,605	96	10,608	95
1987-88	10,977	94	10,968	93
1988-89	9,985	93.5	9,995	92.8

The department is recommending that the standard be raised from 370 to 400 for the eleventh grade class that will take the test in February, 1991. This is the first use of the 1990 Edition of these tests at grade 11. The expectation is that in its initial application, up to sixteen percent of the students taking the test for the first time will fail. If the data from the past four years for the 1984 and 1986 editions of the test, presented in Table 2, are representative, the percentage of failures will be less than sixteen percent.

It is expected that the instructional systems already in place in the districts are sufficient to support the new standards and that the percentage of failures will decrease as the new standards are incorporated into district programs and students approach these tests with increased motivation.